STATISTICS IN MEDICINE
Statist. Med. **18**, 2377–2399 (1999)

ISSUES IN THE STATISTICAL ANALYSIS OF SMALL AREA HEALTH DATA

JON WAKEFIELD* AND PAUL ELLIOTT

Small Area Health Statistics Unit, Department of Epidemiology and Public Health, Imperial College School of Medicine, St Mary's Campus, Norfolk Place, London W2 1PG, U.K.

SUMMARY

The availability of geographically indexed health and population data, with advances in computing, geographical information systems and statistical methodology, have opened the way for serious exploration of small area health statistics based on routine data. Such analyses may be used to address specific questions concerning health in relation to sources of pollution, to investigate clustering of disease or for hypothesis generation. We distinguish four types of analysis: disease mapping; geographic correlation studies; the assessment of risk in relation to a prespecified point or line source, and cluster detection and disease clustering. A general framework for the statistical analysis of small area studies will be considered. This framework assumes that populations at risk arise from inhomogeneous Poisson processes. Disease cases are then realizations of a thinned Poisson process where the risk of disease depends on the characteristics of the person, time and spatial location. Difficulties of analysis and interpretation due to data inaccuracies and aggregation will be addressed with particular reference to ecological bias and confounding. The use of errors-in-variables modelling in small area analyses will be discussed. Copyright © 1999 John Wiley & Sons, Ltd.

1. INTRODUCTION

The availability of geographically indexed health and population data, and advances in computing, geographical information systems and statistical methodology, have enabled the realistic investigation of small area variation in disease risk. In this paper we describe an idealized conceptual framework from which small area analyses may be viewed. We then describe how this view must be adjusted when the necessarily incomplete information available is considered. In particular we consider the implications of *inaccuracies* in population, exposure and health data, and the effect of the *scale* at which the data are recorded. We discuss how *errors-in-variables* modelling can, at least in theory, address some of these problems. Specific difficulties that are emphasized include ecological bias and confounding.

We motivate our discussion by considering the aims and use of small area analyses. To achieve this we distinguish between four types of study:

(i) disease mapping;
(ii) geographic correlation studies;

* Correspondence to: Jon Wakefield, Small Area Health Statistics Unit, Department of Epidemiology and Public Health, Imperial College School of Medicine, St Mary's Campus, Norfolk Place, London W2 1PG, U.K.

CCC 0277–6715/99/182377–23$17.50

(iii) the assessment of risk in relation to a point or line source;
(iv) cluster detection and disease clustering.

Disease mapping is carried out to summarize spatial and spatio-temporal variation in risk. This information may be used for simple descriptive purposes, to provide context for further studies, or, by comparing the estimated risk map with an exposure map, to obtain clues as to disease aetiology. *Geographic correlation studies* exploit geographical variations in exposure to environmental variables (such as air pollution) and life-style factors (such as smoking and diet), again in order to gain clues as to disease aetiology. While the statistical models that are used for disease mapping and geographic correlation studies may be similar,[1] the differing aims distinguish them; disease mapping studies are primarily descriptive, while geographic correlation studies are focused on aetiological questions. *Point source* type studies are appropriate when increased risk close to the source is suspected, or where the source is considered to present a potential environmental hazard. The exposure may be related to a point source, for example, a nuclear installation or a radio transmitter, or a linear source, for example, a road or a powerline. In such cases any increased exposure due to the putative source is likely to extend over a small region and only a highly-localized study will have sufficient geographic resolution to provide an estimate of the associated risk. When a well-defined biological hypothesis is driving the investigation then the interpretation of the results is most straightforward, but where the study is carried out because of a media report, or the worries of the local population, interpretation becomes much more difficult because there is no *a priori* hypothesis.

Finally, detection of individual disease '*clusters*' *or general* '*clustering*', with no associated hypothesis, may be attempted but again interpretation is difficult. Surveillance (cluster detection) is carried out to provide early detection of raised incidence of disease when there is no specific aetiological hypothesis. More general studies of clustering, that is the tendency for disease cases to occur in non-random spatial patterns (allowing appropriately for the underlying population distribution) have a more robust statistical formulation and again may give clues as to aetiology. For example, there is consistent evidence of spatial clustering of Hodgkin's disease (for example, Alexander *et al.*[2]) which, along with other epidemiological evidence and laboratory studies, have suggested a possible infectious aetiology.

We note that the above characterization is convenient for our purposes but there is overlap between the categories. For example, disease mapping may provide information both on individual disease clusters and more generally on clustering, while a point source of exposure may give rise to localized non-random distribution of cases.

We begin our description of an idealized *population/exposure/health outcome* framework within which to carry out small area analyses by stating an obvious fact: individuals are not uniformly distributed in space or time; they are born at a location on a particular date which depends (in probabilistic terms) on the population density on the date, and they then move through space as part of their daily lives or because of migration. During these movements, indexed by time, individuals will travel through numerous exposure surfaces and the integrated exposure will determine the usual biological quantity of interest in a study, the lifelong dose. Individual characteristics such as age, sex and genetic factors, and life-style variables such as smoking and diet, along with lifelong dose due to an exposure of interest, all then contribute to the subsequent disease experience of an individual. Statistical models may be proposed for each of the components of the idealized framework, but suitable data are required for these models to be useful.

The ideal data would consist of precise information on the population of a study region including individual characteristics, movements, personal exposures and subsequent health record. Of course it is never feasible to obtain such information, and a number of simplifications to the model suggested by the idealized framework (assumptions), are imposed by the available data. In many situations the quality of the data and subsequent extent of the assumptions may seriously limit the utility of a study, particularly when one considers that the increases in risk due to many putative exposures are likely to be small. However, from a public health standpoint, there is a often a need to provide a view on a specific question, based on the data at hand, and with careful consideration of the aforementioned shortcomings. At the least this will provide a qualitative answer on whether or not there appears to be any problem. There are a number of examples in which space–time clusters of disease cases have provided clear aetiological information. For example, a cluster of malignant pleural mesothelioma cases in the small Turkish village of Karain was subsequently linked to the identification of exposure to naturally occurring erionite fibres,[3] although this example is atypical in that it relates to an exposure that produced a high excess risk.

We now distinguish between point data and count data. Each of the population, exposure and health data may have associated exact spatial and temporal information (*point* data) or be available as aggregated summaries (*count* data). Point data give the closest link to the conceptual framework but such data are rarely available routinely. Case-control studies provide point data for cases and a set of controls but require epidemiological expertise, to minimize difficulties of selection and other biases, are expensive and time-consuming to carry out, and may not be feasible in given situations. For these reasons case-control studies are not carried out routinely but only when there is sufficient evidence/concern to warrant their use.

Small area studies are complex and we concentrate on particular issues. The outline of this paper is as follows. In Section 2 we review issues relating to population, exposure and health data that are relevant for small area analyses. In Section 3 we consider the problems caused by aggregation and by the inaccuracies that are due to data collection. In particular we discuss how these issues potentially lead to problems of ecological bias and confounding. In Section 4 a statistical framework for small area studies is described and the modelling and analysis of the four types of study, noted above, is placed within the context of this framework. Commonly-made assumptions are also highlighted within the context of this framework. In Section 5 we discuss how errors-in-variables modelling may be utilized in small area studies. Section 6 contains concluding remarks. We do not review the now huge literature on small area methods, instead we highlight key papers at appropriate points. Useful reviews can be found in Marshall,[4] Alexander and Cuzick,[5] Clayton and Bernardinelli,[6] Elliott *et al.*[7] and Alexander and Boyle.[8]

2. DATA ISSUES

As described above, the ideal situation in which to carry out small area studies is when accurately recorded point data are available. Only rarely will we find ourselves in such a situation and in this section we describe the data that are typically available. In the next section we consider the implications of using these data.

2.1. Population data

When a case-control study is carried out, information on the population at risk is obtained via the exact locations of the control sample. More typically, population data, at least for routine

inquiries, are based on aggregated counts. Diamond[9] provides a review of the data that are available for estimating the populations of small areas. National population registers are the gold standard but, as described by Diamond,[9] only rarely are such data available and estimates are typically based on vital registration (births and deaths) and censuses. The latter provide a snapshot of the population on a specific date, stratified by, at the least, age and sex. The raw census counts are themselves estimates, being subject to miscount, with the more likely error being underenumeration. Often an attempt is made to correct the raw counts, for example the 'Estimating with Confidence' project in the U.K. adjusted the 1991 census population statistics.[10] Such work can provide valuable information on the likely discrepancy between actual and estimated population sizes, as can local registers. In particular errors-in-variables modelling (Section 5) may utilize such information.

For inter-censual years, population counts must be estimated and must take into account not only the usual demographic changes (that is, births/deaths) but also migration. Population projections beyond the most recent census (and perhaps before the earliest usable census) will also often be required for small area studies. The frequent lack of a common geography between censuses introduces further problems when a set of population counts by year is produced. In England and Wales, 70 per cent of the censual Enumeration Districts (EDs) changed between the 1981 and 1991 censuses whilst the geographical units were different again in 1971.

2.2. Exposure data

We will use *exposure data* as an umbrella term for all explanatory variables, that is variables associated with disease status, therefore it includes confounders such as social deprivation, as well as other social and environmental factors. The difficulty in obtaining appropriate measures of exposure in small area studies should not be underestimated. Individual exposure sampling and biological monitoring are both costly and invasive, even where reliable and valid measures to estimate exposure are available. Direct individual exposure data are therefore only available very rarely. Environmental monitoring is also expensive and is likely to give only a partial picture of the true exposure over an area, especially if it is exposure integrated over many years that is of most interest. As an alternative, exposure maps may be produced, using for example interpolation via kriging,[11] and from these, individual or area-level estimates may be derived (see Briggs *et al.*[12] for a recent example). The method of construction of such maps is obviously crucially dependent on the type of exposure, the medium (for example, air/soil/water) and the available data. Exposure levels may be based on one or a number of: monitors, emission modelling and dispersion modelling.

Often (for example, Elliott *et al.*[13]) small area studies use distance from putative source as a surrogate for exposure. Distance may relate to individual residences (point data) or to the population centroid (say) of a small area (count data).

A major difficulty in interpretation is the issue of socio-economic confounding. Sources of pollution tend to be in socio-economically disadvantaged areas, whilst deprivation itself is strongly linked to ill health and health-defining behaviour such as smoking. Failure to account for social deprivation can therefore seriously bias investigation of small area health statistics. Area-level indices of deprivation may be constructed from census variables though this may only provide a very crude adjustment for the underlying variables of interest (for example, individual-level smoking).

2.3. Health data

As with population and exposure data, health data may be available with associated point locations, or as aggregated counts, and will potentially be subject to a number of inaccuracies. For any health event there is always the potential for diagnostic error or misclassification, especially at older ages where diagnostic tests and post-mortem examinations are carried out less frequently than at younger ages. Some events may be captured poorly, if at all, in routine registers, for example, early abortions. For others, such as cancers, case registers may be subject to double counting and under-registration as well as diagnostic inaccuracies. Some assessment of the basic quality of the data is therefore essential to inform their use in small area analyses.

3. IMPLICATIONS OF DATA INADEQUACIES

In this section we distinguish between data inaccuracies, many of which will often be checkable and sometimes correctable, and problems due to the different scales at which the data are held, which is essentially fixed and determines the type of analysis which may be carried out.

3.1. Data inaccuracies

For point data arising from a case-control study many data inaccuracies will be alleviated by a careful design. For example problems due to migration may be removed by careful criteria for inclusion; for example, selected individuals should have been resident at their location for a sufficient period.

For count data let Y_i and E_i denote the observed and expected numbers of cases in area A_i, $i = 1, \ldots, I$. These expected numbers are calculated as $E_i = \sum_{jk} p_{jk} N_{ijk}$ where j indexes an age/sex stratification S_j, k a time stratification T_k, p_{jk} is the probability of disease in (S_j, T_k) and N_{ijk} is the number of individuals in (A_i, S_j, T_k) (for further discussion see Section 4.2). Population data problems (underenumeration, migration) and health data problems (double-counting and under-registration, diagnostic accuracy, migration) are not, in general, spatially and temporally neutral and so ignoring their effect will introduce bias into relative risk estimates. Hence areas at apparently high/low risk may simply be reflecting data anomalies. Problems with population denominators are likely to be most acute for small area studies. These data problems may be checked using separate information or other studies/databases. For example, it may be possible to use local registry information to obtain more accurate population counts, or at least to determine the levels of migration in and out of the study region. This will allow, at the least, a qualification of the order of the potential bias. These processes will be time-consuming, and may not obtain a complete set of error-free cases and populations at risk, but will at least allow a more realistic interpretation of the statistical analysis.

The implications of data inaccuracies can be determined by consideration of the effect on the basic relative risk estimate Y_i/E_i. For example, census underenumeration will lead to under-estimation of E_i and hence an overestimate of relative risk. This problem will be most acute when underenumeration occurs in a stratum with a high disease probability. Double-counting and under-registration of cases clearly lead to over- and under-estimation of risk, respectively.

Migration covers both immigration and emigration and both numerator and denominator may be affected. Consequently the overall effects of migration are difficult to determine. In larger areas, migration problems are likely to cancel out since individuals will migrate *within* areas. Various scenarios are possible but migration leads to dilution of effect when individuals who have

been in an exposed area for a number of years migrate to an unexposed area and are then registered with the disease at their new location. The modelling of errors in disease classification, in a general epidemiological context, has been considered by Whittemore and Gong.[14] By assuming individuals have a fixed location, and so ignoring local movements, misclassification of exposure occurs which again introduces bias. We postpone a detailed discussion of the effect of inaccurately-measured exposures to Section 5.

3.2. Scale problems

As described in Section 2, data on cases, populations at risk and exposure variables are typically collected at different levels of spatial aggregation. As an example consider a study in the north-west of England investigating the association between water constituents, in particular magnesium and calcium, and heart disease mortality using routine health data.[15] Cases are available at the postcode level and populations at the ED level. Postcodes contain on average 14 households and EDs on average 400 individuals. Exposure to water constituents is determined through levels in the household supply, measured at water company defined water zones that contain, on average, 45 EDs. Adjustment for deprivation is carried out using the Carstairs' index[16] that is based on census variables and is hence an ED-level variable. Figure 1 shows a subregion of this study and highlights the different geography.

A crude statistical analysis would take the coarsest level of aggregation as the unit of analysis, thus removing the scale problem, but discarding possibly important information.

A major problem in small area studies is the possibility of *ecological bias* which arises when individual response/exposure relationships are estimated from data aggregated across groups. In the spatial context these groups are areal units. The problem arises because only under very strict conditions will a relationship at the individual level remain unchanged when estimated from aggregated data.[17] Ecological bias tends to decrease in magnitude the closer one gets to individual data, that is, as the aggregation areas reduce in size. Ecological bias leads to the ecological fallacy and has generated much discussion in the epidemiology literature; see for example Richardson *et al.*,[18] Piantadosi *et al.*[19] and Greenland and Robin.[20] Ecological bias may arise due to confounding both within and between areas. If these confounders are measured then bias may potentially be reduced but in practice one never knows whether all such confounders have been adjusted for.

4. STATISTICAL FRAMEWORK

In this section we describe the general Poisson process framework within which small area studies may be viewed. This framework has been considered by a number of authors including Diggle[21] and Diggle and Elliott.[22] The latter considered ecological bias and different scales of measurement from this viewpoint.

In Diggle[21] the modelling of risk in relation to a putative source was considered when point data for a sample of controls and disease cases were available. He considered the two-dimensional locations of residence x of controls and cases as arising from inhomogeneous Poisson intensity functions $\lambda_2(x)$ and $\lambda_1(x)$, respectively. This is useful but a major assumption, namely that exposure is associated with a specific location, has been made.

Here we extend the framework to consider, in addition to space, individual characteristics and time. The modelling of the effect of non-spatial variables has been considered previously by

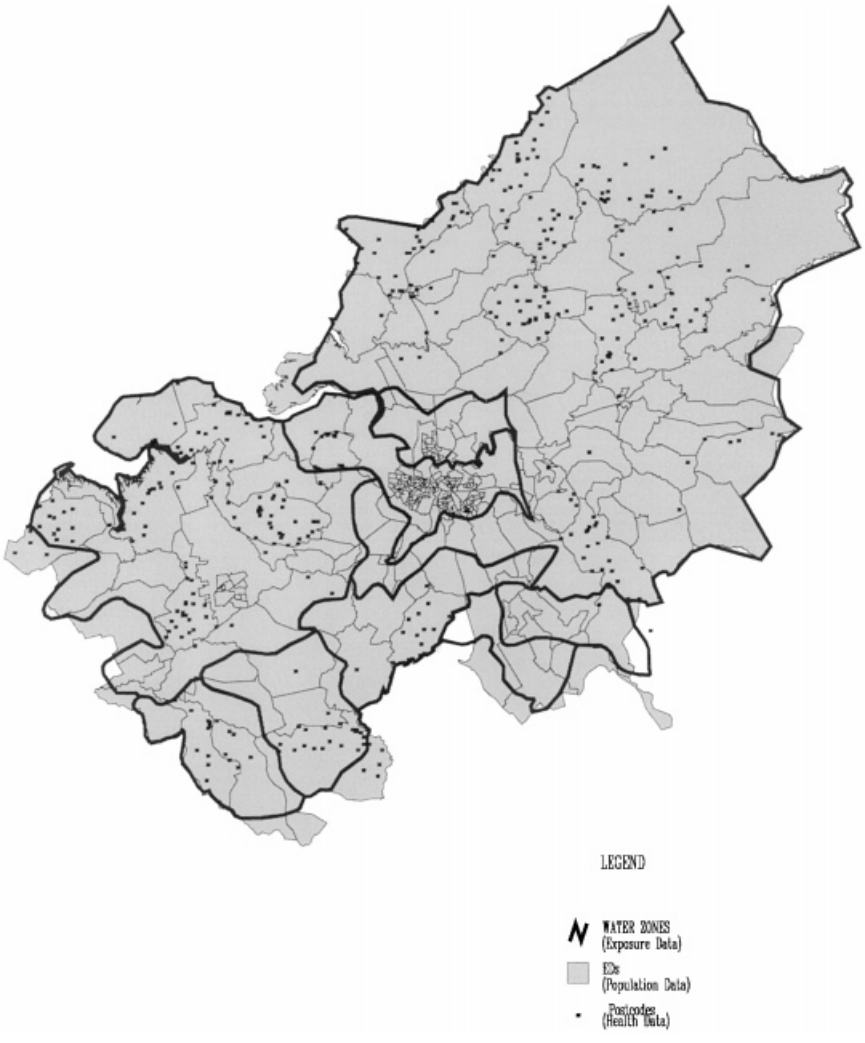

Figure 1. Region of north-west England showing postcodes (only those postcodes lying in geographically large enumeration districts are shown for clarity), enumeration districts and water zones at which, respectively, health, population and exposure data are measured

a number of authors (for example, Diggle and Rowlingson[23]), see Sections 4.1 and 4.2. Let z represent individual characteristics (for example, age, sex, genetic factors and ethnicity) and t calendar time. Then the population at risk, at a location x, with characteristics z and at time t, may be modelled by a Poisson process with intensity function $\lambda_0(x, z, t)$. Similarly, for a non-infectious disease (that is, assuming no interaction between cases), cases are generated with intensity function $\lambda_1(x, z, t)$ where

$$\lambda_1(x, z, t) = \lambda_0(x, z, t) \times p(x, z, t) \tag{1}$$

and $p(x, z, t) = \Pr(\text{disease}|\text{location } x, \text{characteristics } z, \text{time } t)$. Hence cases of the disease form a *thinned* Poisson process. In general, epidemiological investigations aim to learn about $p(x, z, t)$ (place/person/time) and spatial epidemiological investigations are particularly interested in the way in which $p(x, z, t)$ varies with x.

We now consider the analysis of point and count data within this framework, with reference to the four types of study outlined in Section 1.

4.1. Point data

We first consider point, that is, case-control, data since this provides the closest link with the Poisson process framework. Suppose that we obtain a set of cases with locations x_i, $i = 1, \ldots, n$ and a sample of controls with locations x_i, $i = n + 1, \ldots, n + m$ from some region of interest R. The controls may be a random sample from the population at risk, or be cases of another disease which has a similar age/sex profile.

Since we do not, in general, obtain the complete set of non-cases, we are required to introduce further notation. Let $\lambda_2(x, z, t)$ denote the intensity function of the controls where

$$\lambda_2(x, z, t) = c \times [\lambda_0(x, z, t) - \lambda_1(x, z, t)] \tag{2}$$

and $0 < c \leqslant 1$ represents the probability of selection for a non-case. The selection mechanism implies the quantity c although this is rarely known (and is not required to be known for inference concerning spatial location, see below). The reason that c is not known is that in practice we do not select stochastically from the pool of controls. Suppose we require three times as many controls as cases. We would first identify the n cases and then simply select $m = 3 \times n$ non-cases from a convenient source. For example, in a study of asthma and proximity to roads via hospital admissions, non-respiratory admissions may be used as the pool of controls. In such a study car accident admissions would be excluded to avoid selection bias.

For completeness we may also define the intensity function of the *selected* cases:

$$\lambda_3(x, z, t) = d \times \lambda_1(x, z, t)$$

where $0 < d < 1$ represents the probability of a case being registered. Hence if we have under-registration (for example) over all covariate groups in a particular period (t_1, t_2) then we will have $d < 1$ for $t \in (t_1, t_2)$.

Now suppose we condition on the *locations* of the cases and controls and let Y_i' represent a Bernoulli random variable with $Y_i' = 1$ for a case and $Y_i' = 0$ for a control. We denote by $p'(x, z, t)$ the probability of being a case. Then, using the standard argument for inference for case-control data via logistic regression, we have

$$p'(x, z, t) = \frac{d \times p(x, z, t)}{d \times p(x, z, t) + c \times [1 - p(x, z, t)]}$$

or, equivalently

$$p'(x, z, t) = \frac{\lambda_3(x, z, t)}{\lambda_3(x, z, t) + \lambda_2(x, z, t)}$$

both of which lead to

$$\frac{p'(x,z,t)}{1-p'(x,z,t)} = \frac{\lambda_3(x,z,t)}{\lambda_2(x,z,t)} = \frac{p(x,z,t)}{1-p(x,z,t)} \times \frac{d}{c}$$

providing a link between $p'(x,z,t)$ and $p(x,z,t)$. Hence, although we obviously cannot recover $\lambda_3(x,z,t)$ and $\lambda_2(x,z,t)$, we can estimate the ratio of the odds of disease at different locations.

4.1.1. Disease mapping

Disease mapping studies are most frequently associated with count data, but if we use as a definition the estimation of a geographic risk surface then such studies may be carried out using point data also.

We first consider methods based on kernel density estimation. Such methods are non-parametric and so it is not possible to model explicitly the effects of individual covariates z or time t, using the basic approach (see below for an extension). Obviously risk estimates may be constructed for different covariate groups and time periods, though sparsity of data restricts the number of such maps that may be produced. To simplify notation we, for the moment, suppress dependence on z and t, and let $\lambda_3(x)$ and $\lambda_2(x)$ denote the intensity functions of cases and controls, respectively (where it is assumed that the controls are a simple random sample from the non-cases). Then the cases and controls may be viewed as random samples of sizes n and m from the probability density functions $\lambda_3^*(x)$ and $\lambda_2^*(x)$, respectively, where

$$\lambda_j^*(x) = \frac{\lambda_j(x)}{\int_R \lambda_j(x)\,\mathrm{d}x}$$

for $j = 2, 3$. Bithell,[24] Lawson and Williams[25] and Kelsall and Diggle[26,27] then used kernel density estimation to provide an estimate of the *relative risk surface* $\lambda_3^*(x)/\lambda_2^*(x)$. Strictly speaking this surface is proportional to the odds of, and not the risk of, disease, but for rare diseases these quantities are approximately equal. These techniques allow the direct visualization of the intensity functions $\lambda_3(x)$ and $\lambda_2(x)$ though it is the ratio of these that is of interest.

The effects of covariates z and time t may be accounted for using as a control a sample of cases from a disease with a similar pattern of disease over z and t (for example, lung and larynx cancer in relation to smoking[21]).

More recently, Kelsall and Diggle[28] have proposed an alternative method using a generalized additive model.[29] The latter may be used explicitly to model the odds of disease as a function of non-spatial variables and spatial location. Specifically the logistic generalized additive model

$$\text{logit}\, p'(x,z,t) = \alpha + f_z(z) + f_t(t) + f_x(x)$$

is used, where f_z and f_t are linear functions of unknown parameters, and $f_x(x)$ is estimated using kernel regression.

4.1.2. Geographic correlation studies

Geographic correlation studies in general estimate associations between aggregated population/exposure and health data, and hence utilize count data. However, in some instances (for

example, case-control studies) individual health and population point data may be available but exposure may be measured at a larger scale. For example, the same exposure score may be assigned to all individuals within a certain distance of a pollution monitor. In this case standard logistic regression approaches to the analysis of case-control data may be used, that is

$$\text{logit}\, p'(x, z, t) = \alpha + \beta z + \gamma t + \delta W(x) \tag{3}$$

where $W(x)$ denotes the pollution at location x. There is still an ecological level to the analysis, namely the level at which the exposure is measured.

4.1.3. *Point/line sources*

The effect on the odds of disease of a point/line source may be investigated via standard logistic regression using a direct measure of exposure (as in (3)). Alternatively a surrogate for exposure, such as distance from pollution source (for example, Cook-Mozzafari *et al.*[30]), may be used. A disadvantage of this approach is that as distance tends to infinity the probability of disease tends to zero and not baseline, as required.

The model given by (3) may also be used when each individual in the study is assigned a unique pollution score $W(x)$ from a pollution map, thus removing the ecological aspect. For example the Small-Area Variations In Air Quality and Health (SAVIAH) case-control study[12] investigated the relationship between childhood wheeze, levels of nitrogen dioxide and individual-level explanatory variables. Nitrogen dioxide measures were available at a number of monitor sites and these were modelled as a function of location-specific variables such as altitude, traffic volume and land cover. These latter variables were also available at each of the case and control locations and so approximate nitrogen dioxide levels $W(x)$ could be calculated for each of the study individuals.

To more realistically model the exposure/risk relationship, Diggle and Rowlingson[23] proposed the model

$$\text{logit}\, p'(x, z, t) = \alpha + \beta z + \gamma t + \log f(|x - x_0|, \theta) \tag{4}$$

where $f(|x - x_0|, \theta)$ represents a simple function, depending on parameters θ, to describe the effect of being at location x, relative to the location of the point source x_0. This function was taken to be a simple monotonic function of distance, which tends to one as distance tends to infinity.

Diggle[21] did not condition on locations and assumed that

$$\lambda_1(x) = \rho \lambda_2(x) f(|x - x_0|, \theta)$$

where ρ is a scaling parameter and the 'nuisance parameter' $\lambda_2(x)$ was estimated using kernel density estimation. Unfortunately this method is sensitive to the choice of smoothing parameter and, since $\lambda_2(x)$ is not of interest, an approach using (4) is preferable.

4.1.4. *Cluster detection and disease clustering*

A large number of methods have been proposed to detect individual clusters or general clustering using individual-level data.[5,8,31] Cluster detection, or surveillance, may be carried out using the kernel-type methods described in Section 4.1.1, though the statistical properties (for example, sensitivity/specificity) of such an approach are unknown. In particular the choice of smoothing parameter is likely to be crucial.

When a particular collection of cases is alleged to be a cluster, then any investigation should define a population of interest z_c and a time period of interest t_c. Often these parameters are selected *a posteriori* leading to the problem of 'boundary shrinkage'. The smaller the population at risk and the shorter the time period that contains the cases within the 'cluster', the larger the apparent excess in risk. The statistical problem is to estimate the odds surface $\lambda_3(x, z, t)/\lambda_2(x, z, t)$ for $z \in z_c$ and $t \in t_c$. Locations x may then be highlighted as the location of a 'cluster' if they exceed an epidemiologically significant risk r_c. In this way an observed collection of cases has been replaced by an estimated risk surface. As Alexander and Cuzick[5] point out, a mathematical definition of a cluster or clustering is far easier to determine than a definition based on an empirical set of data.

Three related cluster detection methods are described by Openshaw *et al.*,[32] Turnbull *et al.*[33] and Besag and Newell.[34] Each of these methods considers the number of cases located within circles drawn about specific points. The observed significance level of each configuration is then assessed via Poisson tail probabilities using the populations within each circle. No explicit use of the framework described above is used; the methods are essentially non-parametric and vary in the degree of arbitrariness in the way in which the required circle radii are defined, and in the rigour with which the multiple testing problem is acknowledged.

Cuzick and Edwards[35] examine general clustering by, for each case, counting the number of cases within the k nearest neighbours. An overall level of significance is produced but no estimate of the geographic scale of clustering. A method which gives such information is that of Diggle and Chetwynd[36] who estimate the difference between the clustering in the underlying case and control spatial processes. Extensions to these basic methods may be found in Alexander and Boyle.[8]

A general problem with clustering is that if it is due to a specific exposure variable then it is extremely unlikely that the scale will be constant across the study region. This may be assessed by examining estimated risk surfaces produced by kernel methods.

4.2 Count data

Suppose the study region R is divided into a set of areal units A_i, $i = 1, \ldots, I$ (for example, EDs). Suppose also that we obtain the data stratified by a set of individual groupings $S_j, j = 1, \ldots, J$ (for example, five-year age bands and sex) and by time intervals T_k, $k = 1, \ldots, K$ (for example, calendar years). The available data then typically consist of:

(i) Y_{ijk}, the number of cases in area A_i within stratum S_j and time period T_k.
(ii) N_{ijk}, the population at risk in area A_i within stratum S_j and time period T_k. We assume that N_{ijk} includes both cases and non-cases which would be true if the counts were obtained from the census.
(iii) W_{ik}, the values of explanatory variables measured for area A_i in time period T_k. These variables may include characteristics such as deprivation index, pollution scores which may be measured at particular locations within area A_i, or 'hospital effects', due for example to differences in diagnosis or coding between hospitals, in which case many neighbouring areas will receive the same score.

For simplicity of notation we have imposed a common geographical/temporal scale on the cases, populations and area characteristics. We note that the stratification by calendar year, age and sex can in principle be taken as fine as possible, for example, annual 5-year age–sex counts. The grouping into areas is not flexible.

The assumption of a Poisson process then implies that $N_{ijk} \sim$ Poisson (λ_{0ijk}) and $Y_{ijk} \sim$ Poisson(λ_{1ijk}) where

$$\lambda_{lijk} = \int_{A_i} \int_{S_j} \int_{T_k} \lambda_l(x, z, t) \mathrm{d}t \, \mathrm{d}z \, \mathrm{d}x$$

for $l = 0, 1$. Note that with count data we can gain no information within the cell (A_i, S_j, T_k). In particular we do not know the form of the relationship within a class and so we cannot evaluate these integrals.

To model $p(x, z, t)$ we are required to make further assumptions. In particular we may take $p(x, z, t) = p(z, t) \times p(x|z, t) = p(z, t) \times p(x|t)$. This assumption states that the risk decomposes into a person/time component and a spatial effect which depends on time. For example, the probability of disease at a particular point may change at the time at which a pollution source become active (allowing for a suitable lag period).

An assumption which is then implicitly made is that the discretization (S_j, T_k) is chosen so that $p(z, t)$ is approximately constant over stratum S_j and time period T_k, that is, $p(z, t) \approx p_{jk}$ for $z \in S_j$ and $t \in T_k$. This would obviously not be true if the stratification over z were taken to be too coarse and also if very large time periods were considered. Bias will result if each of these is far from the truth.

The quantity p_{jk} may be estimated in a number of ways. The relationship may be assumed to follow some mathematical form, for example a log-linear model whose parameters may then be estimated. More commonly the p_{jk} are estimated using either a larger region within which the study lies (external standardization) or from all of the data in the study region (internal standardization), that is $\hat{p}_{jk} = Y_{+jk}/N_{+jk}$. This shows the difficulty of taking too fine stratification and too narrow time periods – for rare diseases there will be insufficient data for precise estimates to be obtained. We note that if inappropriate rates are used then bias will also result. For example, for many diseases there are large differences between the south and north of England and so the use of national rates for local studies may not be appropriate.

We now assume

$$p(x|t) = f(W(x, t), U(x, t))$$

where we have covariates $W(x, t)$, for which there are available measurements, and *unmeasured* covariates (which we assume influence risk) $U(x, t)$.

We then have

$$\begin{aligned} \lambda_{1ijk} &= \int_{A_i} \int_{S_j} \int_{T_k} \lambda_0(x, z, t) \times p(z, t) \times p(x|t) \mathrm{d}t \, \mathrm{d}z \, \mathrm{d}x \\ &= p_{jk} \int_{A_i} \int_{S_j} \int_{T_k} \lambda_0(x, z, t) \times f(W(x, t), U(x, t)) \mathrm{d}t \, \mathrm{d}z \, \mathrm{d}x. \end{aligned} \tag{5}$$

If we assume $W(x, t) = W_{ik}$ and $U(x, t) = U_{ik}$, for $x \in A_i$, $t \in T_k$, then we have introduced the potential for ecological bias (since the covariates are assumed to be constant) but we have

$$\begin{aligned} \lambda_{1ijk} &= p_{jk} \times f(W_{ik}, U_{ik}) \times \int_{A_i} \int_{S_j} \int_{T_k} \lambda_0(x, z, t) \mathrm{d}t \, \mathrm{d}z \, \mathrm{d}x \\ &= p_{jk} \times f(W_{ik}, U_{ik}) \times \lambda_{0ijk}. \end{aligned} \tag{6}$$

Diggle and Elliott[22] given an example of the effects of ecological bias using a point source example. They consider location only, and for a generic area A examine

$$\lambda_1 = \int_A p(x)\lambda_0(x)\,\mathrm{d}x.$$

In this case ecological bias is introduced when we try to estimate $p(x)$ from area level data only. The conditions for no ecological bias are apparent, the functions p and λ_0 should be such that we can write

$$\lambda_1 = |A|^{-1} \int_A p(x)\,\mathrm{d}x \int_A \lambda_0(x)\,\mathrm{d}x$$

which occurs when: (i) $p(x)$ is constant in A; (ii) $\lambda_0(x)$ is constant in A; or (iii) the random variables $p(X)$ and $\lambda_0(X)$, where X is uniformly distributed on A, are uncorrelated. As Diggle and Elliott comment, none of these is likely to be true in practice. From these arguments one possibility would be to deform the geographic area in order to produce constant populations at risk (that is, λ_0 constant) but such approaches have so far proved unsuccessful.[37]

If we condition upon the population totals we have

$$Y_{ijk}|N_{ijk} \sim \text{Binomial}\ (N_{ijk}, p_{ijk})$$

where

$$\begin{aligned} p_{ijk} &= \frac{\lambda_{1ijk}}{\lambda_{0ijk}} \\ &= p_{jk} \times f(W_{ik}, U_{ik}). \end{aligned} \tag{7}$$

Hence λ_{0ijk} is effectively estimated by N_{ijk}. In the case of rare diseases, Y_{ijk} is approximately Poisson $(N_{ijk}p_{ijk})$.

4.2.1. Disease mapping

Referring to the above framework, recall that the U_{ik} are unobserved covariates. We now make the natural modelling assumption that

$$\begin{aligned} f(W_{ik}, U_{ik}) &= g(W_{ik}, \beta) \times \exp(U_{ik}\psi) \\ &= g(W_{ik}, \beta) \times \rho \times \eta_{ik} \end{aligned} \tag{8}$$

where the η_{ik} are such that $E[\eta_{ik}] = 1$ and may be viewed as *random effects* which account for unmeasured covariates. Then, from (7) and (8)

$$Y_{ijk} \sim \text{Poisson}(E_{ijk} \times g(W_{ik}, \beta) \times \rho \times \eta_{ik})$$

where the *expected numbers* $E_{ijk} = N_{ijk}p_{jk}$. Summing over all strata and time periods, and assuming $W_{ik} = W_i$ and $\eta_{ik} = \eta_i$ for all k, we obtain the familiar *disease mapping* model[38]

$$Y_i \sim \text{Poisson}(E_i \times g(W_i, \beta) \times \rho \times \eta_i)$$

where $Y_i = \sum_j \sum_k Y_{ijk}$ and $E_i = \sum_j \sum_k N_{ijk}p_{jk}$. Frequently $g(W_i, \beta) = \exp(\beta W_i)$.

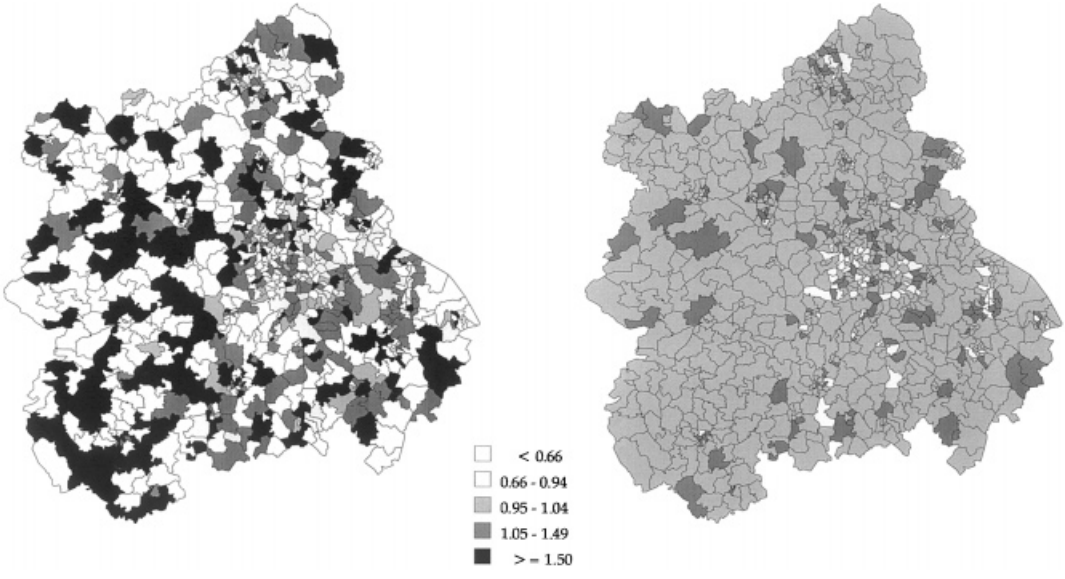

Figure 2. Age, sex and deprivation adjusted relative risks of cancers of the brain and central nervous system for electoral wards in West Midlands region, England. Unsmoothed and smoothed risk estimates are shown on the left and right, respectively (from Eaton *et al.*[39] with permission)

Mapping the raw rates Y_i/E_i is notoriously dangerous (for example, Clayton and Kaldor[38]) since sparsely populated areas have estimates with large standard errors and hence produce unstable rates. On the other hand, mapping the Poisson significance levels highlights areas with large populations, since these are more likely to produce significant results. The left hand panel of Figure 2 shows a map of 'unsmoothed', that is, raw estimates (adjusted for age, sex and deprivation within the expected numbers) of brain cancer incidence for 1974–1986 across electoral wards in the West Midlands region of England.[39] We see that the large, sparsely populated rural areas tend to have more extreme rates which might cause concern to local residents/health officials.

To 'smooth' the rates the η_i may be assumed to be independent realizations from some distribution, for example the gamma or log-normal. Hence *heterogeneity* of rates is modelled. The right hand panel of Figure 2 shows smoothed estimates of risk based on a gamma model for the rates. We see that a greater degree of smoothing has occurred in those rural areas with smaller populations. There do not now appear to be any areas with excessively large rates.

Alternatively the η_i may be assumed to be spatially correlated to give a *clustering* model. In this case a form of conditional autoregressive (CAR) model is a common choice (for example, Besag and Kooperberg[40]). Besag *et al.*[41] specified a model which allowed for both heterogeneity (via spatially-independent random effects) and clustering (via spatially-dependent random effects).

Recent extensions to this basic framework include Best *et al.* (Best, Ickstadt and Wolpert, 1998, in preparation) who apply the gamma random field model of Wolpert and Ickstadt[42] and Kelsall and Wakefield (1998, submitted for publication). The latter assumes that the underlying risk follows a Gaussian process and hence avoids the neighbourhood structures of commonly-used

disease models (for example, the 'common boundary' specification) that are rarely realistic when the areas are of different shapes and sizes.

4.2.2. Geographic correlation studies

Richardson[17] provides a review of methods for geographic correlation studies. These range from simple graphical comparison of maps of exposure and estimated rates, to Poisson regression techniques of the form described in the previous section, and hence relate explicitly to the Poisson process framework. One of the important considerations is to acknowledge the spatial correlation that is present in the data.[1] Wakefield and Morris[15] discuss this and other issues (including overdispersion) in the context of the study mentioned in Section 3.2 investigating the relationship between heart disease mortality and magnesium and calcium in water. In Section 5.3 we discuss an *errors-in-variables* approach to the modelling of the relationship between exposure and disease.

4.2.3. Point/line sources

A simple semi-parametric method for investigating increased risk close to a point source is due to Stone.[43] A Poisson model $Y_i \sim \text{Poisson}(E_i\mu_i)$ is assumed where the labels $i = 1, \ldots, I$ are such that the area centroids have been ranked in increasing distance from the point source. The null hypothesis $H_0 : \mu_1 = \cdots = \mu_I = \mu$ is then tested versus $H_1 : \mu_1 \geqslant \mu_2 \geqslant \cdots \geqslant \mu_I$. The unconditional version of the test takes $\mu = 1$ but has the undesirable property that H_0 may be rejected simply because the risk in the study region as a whole is elevated or reduced. The conditional version with μ unspecified is therefore generally preferred. There are a number of difficulties with Stone's test, in particular no modelled risk as a function of distance is produced. Such a function, with associated standard errors, is a highly informative summary. In the original formulation, covariates could only be dealt through the expected numbers; the test has now been extended by Morton-Jones *et al.*[44] Lawson[45] considers radial and directional health effects associated with a point source.

Diggle *et al.*[46] proposed a framework for investigating disease risk close to a putative source when only count data are available. The model was of the form

$$Y_i \sim \text{Poisson}(E_i\mu_i)$$

where

$$\mu_i = \rho \times \exp(\beta W_{i1}) \times f(W_{i2}, \theta)$$

where $W_{i2} = |x_i - x_0|$ represents the distance of the centroid of ED i from the location of the point source x_0, and W_{i1} represents area-level covariates (for example, deprivation). The function $f(\cdot, \cdot)$ is the same as that specified in equation (4). Their formulation did not explicitly include random effects though overdispersion was incorporated; maximum likelihood was used for estimation. Wakefield and Morris (1998, submitted for publication) consider a Bayesian version and explicitly embed the above model within a disease mapping framework.

4.2.4. Cluster detection and disease clustering

As with point data, a large number of approaches have been proposed for cluster detection and disease clustering with count data. A simple test for heterogeneity of rates is due to Potthoff and

Wittinghill.[47,48] This method and its application to clustering is discussed further in Muirhead and Butland.[49]

A number of the methods outlined in Section 4.1.4 can be applied to count data. As we commented in Section 1, the disease mapping models of Section 4.2.1, with priors modelling spatial dependence, may be used to estimate the size of clustering between 'neighbouring' areas.

5. ERRORS-IN-VARIABLES

We use the term *errors-in-variables* to describe any non-response variable which is measured inaccurately. As detailed in Section 2, in the context of small area studies we must consider errors in population data and the use of inaccurate disease rates, which leads to non-exact expected numbers, and error in data on explanatory variables. We first consider the general effect of errors in both expected numbers and explanatory variables before specific types of error are considered. We finally consider a general Bayesian framework within which errors-in-variables modelling may be carried out. A general description of errors-in-variables modelling may be found in Fuller[50] and Carroll *et al.*[51] Richardson[52] and Wakefield and Stephens[53] provide reviews of errors-in-variables modelling from a Bayesian perspective.

5.1. Errors in expected numbers

In the following we consider count data only and a single area A (we therefore drop the subscript i). Let E^t denote the notional *true* expected disease count for a particular area and E^o the *observed* (or estimated) value. Recall that $E^o = \sum_j \sum_k p_{jk} N_{jk}$ where S_j represents an age/sex stratification, T_k a time period and N_{jk} is the estimated number of individuals in stratum S_j, time period T_k and in area A. The error in E^o may be due to errors in the estimation of population counts and/or errors in the estimated probabilities p_{jk} (for example, because they are based on small numbers). Within the Poisson framework we have $E[Y|E^t] = E^t\mu$ but if instead the estimated expected numbers are used we have

$$E[Y|E^o] = E[E^t|E^o]\mu \tag{9}$$

and

$$\text{var}(Y|E^o) = E[E^t|E^o]\mu + \text{var}(E^t|E^o)\mu^2. \tag{10}$$

Hence we see that we recover the same form of mean model but the variance in the observed data is inflated via the second term in (10). In general the use of a non-exact expected number will introduce bias into the estimation of μ, confidence intervals will be too narrow and test procedures will have poor properties.

5.2. Errors in exposure variables

To see the effect of using variables measured with error let W^t denote the *true* value of an explanatory variable and W^o the *observed* (or surrogate) value. Suppose we have a response Y related to W^t via

$$E[Y|W^t] = \mu(\beta; W^t)$$

and

$$\text{var}(Y|W^t) = v(\phi; W^t).$$

Then the use of W^o gives

$$\begin{aligned} E[Y|W^o] &= E[\mu(\beta; W^t)|W^o] \\ &\neq \mu(\beta; E[W^t|W^o]) \end{aligned} \tag{11}$$

in general. Only when $\mu(\beta; W^t)$ is linear do we obtain equality. Turning now to the variance we have

$$\text{var}(Y|W^o) = E[v(\phi; W^t)|W^o] + \text{var}(\mu(\beta; W^t)|W^o). \tag{12}$$

Thus the variability in the observed data is inflated by the second term. As an example for the log-linear model $Y \sim \text{Poisson}(E^t \exp(\beta_0 + \beta W^t))$, $E[Y|W^o]$ is no longer a log-linear model and we will see extra-Poisson variability.

In general the effects of errors-in-variables will depend on the type of error which is being introduced into the explanatory variable. We now consider some examples.

5.3. Types of error

In a spatial context there are few examples of errors-in-variables modelling. Jordan *et al.*[54] is a recent example of an ecological study and Bernardinelli *et al.*[55] an example of a disease mapping study.

In the context considered here there are a number of types of errors-in-variables, in particular:

(i) variables subject to measurement error;
(ii) interpolated variables; and
(iii) variables only available as an area-level measure.

5.3.1. Measurement error

Measurement error includes errors introduced by the instrument used to measure the variable. In small area applications these may occur when environmental exposure are measured in air, soil and water. In many instances it will be straightforward to model these errors since information will be available on the measurement process. Another situation in which measurement error modelling may be used is when an exposure score is derived from a deterministic formula. For example, in a study of the relationship between powerlines and brain cancers, the electromagnetic field strength at particular locations may be derived from the voltage and load of the line and the distance of residence from the line. This information is combined using a deterministic formula but some idea of the uncertainty in the resultant field strength is available and may be utilized within the errors-in-variables model.

Measurement error models may also be used for lifestyle variables which are measured on individual study participants, for example dietary variables.

As described in Section 4.1.3, a common approach to modelling the association between disease risk and the location of a point or line source is to assume a simple relationship between

risk and distance. Distances from putative sources that are used as surrogate exposures may be extracted using geographic information systems but will not be exact, particularly for small area centroids, but will have an associated precision (for example, within $\pm$ 10 m).

5.3.2. *Interpolated variables*

Another errors-in-variables situation arises when the spatially varying exposure quantity is observed at a finite set of locations (for example, an air pollution monitoring network) and we wish to *interpolate* the value at additional locations. Data of these kind are often referred to as 'geostatistical' data with kriging being a traditional solution, see for example Cressie[11] and Diggle *et al.*[56]

For point health data we require the value $W(x)$ of the exposure at the location x of the study individual, given the values at N points, $W(x_1), \ldots, W(x_N)$. Viewing these as observations from an unknown spatial process, a predicted mean and variance for the random variable $W(x)$ may be derived. These may then be used directly within an errors-in-variable model. Alternatively a joint model may be specified for the stochastic process $W(x)$ and the risk surface $p(x)|W(x)$ where $p(x)$ is the probability of disease at location x. Currie[57] considers the effect of replacing the true exposure $W(x)$ via a kriging estimate $\hat{W}(x)$.

For count health data the average exposure in an areal unit may be estimated using a similar procedure. Measurement error may also be incorporated in the observations $W(x_j), j = 1, \ldots, N$.

5.3.3. *Area-level measures*

As discussed in Section 2, a number of important explanatory variables may only be measured at the area level, for example deprivation. Wakefield and Stephens[53] describe an analysis in which deprivation was modelled via an errors-in-variables approach. We may also obtain an area-level measure of an exposure, for example, ambient air pollution may be measured. In this case if we need an exposure measure for an individual within the area then a *Berkson* error-in-variables model (for example, Fuller[50]) may be appropriate (see below).

5.4. Multi-level framework

In this section we consider a three-stage Bayesian model which is appropriate for count data with errors-in-variables. The model for point data has a similar structure though in this case expected numbers are not relevant.

Stage 1. Here we model the response Y as a function of the true expected numbers and explanatory variables. We have

$$Y \sim \text{Poisson}(E^t \mu(\beta; W^t))$$

where β represents a vector of parameters. We denote the probability distribution of Y as $p(Y|E^t, W^t, \beta)$. Throughout p will denote a generic probability distribution.

Stage 2. At this stage we model the errors in the realized expected numbers and explanatory variables. We assume independence between these two random quantities. For a conventional errors-in-variable models we have $p(E^o|E^t, \phi_E)$ and $p(W^o|W^t, \phi_W)$ where ϕ_E and ϕ_W denote parameters that model the relationship between the observed and true variables. A Berkson error model for W^o would instead specify $p(W^t|W^o, \phi_W)$.

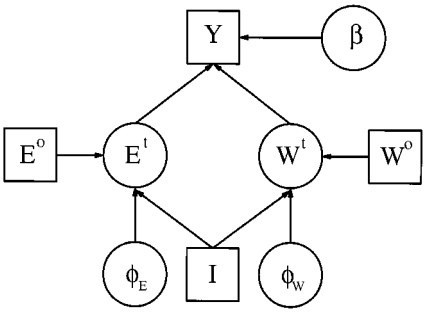

Figure 3. Graphical model representation of the Poisson model with errors-in-variables in the expected numbers E and explanatory variables W, superscript o denotes 'observed' and t 'true'

At this stage we also specify a prior distribution for β.

Stage 3. For the conventional errors-in-variables model we would specify $p(E^t|I)$ and $p(W^t|I)$ where I represents additional information which may be used to inform the priors. For example, for modelling the E^t one my have local information on the size of the populations or some information concerning migration patterns. For Berkson errors for W^o we do not require a prior for W^t.

We also specify prior distribution for ϕ_E and ϕ_W at this stage.

Richardson and Gilks[58] consider errors-in-variables models in epidemiology. Following Clayton[59] they refer to $p(Y|E^t, W^t, \beta)$ as the *disease model*, $p(E^o|E^t, \phi_E)$, $p(W^o|W^t, \phi_W)$ as the *measurement models* and $p(E^t|I)$, $p(W^t|I)$ as the *exposure models*. Figure 3 gives a graphical model representation[60] of the structure in the case of the classical (that is, not Berkson) errors-in-variables model.

With the conditional independence assumptions that are implied by Figure 3 (for example, conditional on the true explanatory variable W^t, the response Y is independent of the observed explanatory variable W^o), we obtain the following posterior for all of the unobservable quantities:

$$\begin{aligned} p(\beta, \phi_E, \phi_W, E^t, W^t|Y, E^o, W^o, I) = c \times p(Y|E^t, W^t, \beta) \\ \times p(E^o|E^t, \phi_E)p(W^o|W^t, \phi_W) \\ \times p(E^t|I)p(W^t|I)p(\beta)p(\phi_E)p(\phi_W) \end{aligned} \quad (13)$$

where $c^{-1} = p(Y, E^o, W^o|I)$ is the normalizing constant. The multi-level model described above is easily implemented using Markov chain Monte Carlo.[58,61]

There are a number of difficulties associated with errors-in-variables modelling. First we must be able to make appropriate distributional assumptions. The most difficult of these are $p(E^o|E^t, \phi_E)$ and $p(W^o|W^t, \phi_W)$ since in small area applications we are rarely able to obtain a validation data set, that is, a data set in which both the observed and true expected numbers/explanatory variables are available. The sensitivity of conclusions to distributional assumptions is therefore an important step in the analysis. Recent errors-in-variables work has concentrated on semi-parametric approaches to remove this dependence.[62] A second difficulty is that the above general formulation offers little insight into the effect of measurement error; only in specific circumstances are analytical results available.

Finally, errors-in-variables modelling should not be viewed as an approach by which poor data can be turned into 'good data'. A good design remains of paramount importance. Problems of ecological bias cannot be solved using errors-in-variables modelling. It would also be difficult to turn exposure at a location into personal exposure, that is, to account for the fact that a person does not stay at their residence for 24 hours a day.

These difficulties notwithstanding, we believe that errors-in-variables modelling is an exciting, and currently under-used, approach in small area modelling.

6. CONCLUDING REMARKS

The effects of low levels of environmental pollution on health are largely unknown but are likely to be much less than the major determinants of disease such as smoking and diet. For example, Doll and Peto[63] attribute 2 per cent of cancer deaths to pollution (the majority of which are due to air pollution). This is a figure which they acknowledge is very approximate due to the lack of investigations. Small area studies are important, therefore, in order to gain a greater understanding of small-scale variability in disease risk, to obtain clues as to disease aetiology, and to address public concerns over specific sources of pollution.

In this paper we have highlighted a number of data deficiencies which make the analysis and interpretation of small-area data difficult. The greatest problem lies in the frequent lack of a good measure of exposure. Without such a measure the detection of a small but important increase in risk is likely to fail. Consequently there is a great need for better exposure data and exposure models. The use of errors-in-variables models in this context is in its infancy but we believe it has a major role to play in the future.

To analyse small-area data successfully it is vital that epidemiologists, statisticians, pollution modellers, geographers, data providers and computer scientists share their expertise. With this collaboration, and with further advances in data collection, computing, geographical information systems and statistical methodology, the challenge is to learn more about the complex, yet important, relationship between individuals at risk, environmental exposure and potential effects on health.

ACKNOWLEDGEMENTS

The authors would like to thank Dr. Julia Kelsall of Lancaster University for detailed comments which greatly improved this article. The Small Area Health Statistics Unit is funded by a grant from the Department of Health, Department of the Environment, the Health and Safety Executive, the Scottish Office, the Welsh Office and the Department of Health and Social Services (Northern Ireland). This work was also supported in part by an equipment grant from the Wellcome Trust (0455051/Z/95/Z). The views expressed in this paper are those of the authors and not necessarily those of the funding departments.

REFERENCES

1. Clayton, D. G., Bernardinelli, L. and Montomoli, C. 'Spatial correlation in ecological analysis', *International Journal of Epidemiology*, **22**, 1193–1202 (1993).
2. Alexander, F. E., Williams, J., Cartwright, R. A. and Ricketts, T. J. 'A specialist leukaemia/lymphoma registry in the UK. Part 2: clustering of Hodgkin's disease', *British Journal of Cancer*, **60**, 948–952 (1989).

3. Baris, Y. I., Simonato, L., Atrinli, M., Pooley, S., Saracci, R., Skidmore, J. and Wagner, C. 'Epidemiological and environmental evidence of the health effects of exposure to erionite fibres: a four year study in the Cappodocian region of Turkey', *International Journal of Cancer*, **39**, 10–17 (1987).
4. Marshall, R. 'A review of methods for the statistical analysis of spatial patterns of disease', *Journal of the Royal Statistical Society*, *Series A*, **154**, 421–441 (1991).
5. Alexander, F. and Cuzick, J. 'Methods for the assessment of disease clusters', in Elliott, P., Cuzick, J., English, D. and Stern, R. (eds), *Geographical and Environmental Epidemiology*: *Methods for Small-area Studies*, Oxford University Press, Oxford, 1992, pp. 238–250.
6. Clayton, D. G. and Bernardinelli, L. 'Bayesian methods for mapping disease risk', in Elliott, P., Cuzick, J., English, D. and Stern, R. (eds), *Geographical and Environmental Epidemiology*: *Methods for Small-area Studies*, Oxford University Press, Oxford, 1992, pp. 205–220.
7. Elliott, P., Martuzzi, M. and Shaddick, G. 'Spatial statistical methods in environmental epidemiology: a critique', *Statistical Methods in Medical Research*, **4**, 149–161 (1995).
8. Alexander, F. E. and Boyle, P. *Methods for Investigating Localised Clustering of Disease*, International Agency for Research on Cancer Scientific Publications, No. 135, 1996.
9. Diamond, I. 'Population counts in small areas', in Elliott, P., Cuzick, J., English, D. and Stern, R. (eds), *Geographical and Environmental Epidemiology*: *Methods for Small-area Studies*, Oxford University Press, Oxford, 1992, pp. 96–105.
10. Simpson, S., Tye, R. and Diamond, I. 'What was the real population of local areas in mid-1991', Estimating with Confidence Project Working Paper 10, 1995.
11. Cressie, N. A. C. *Statistics for Spatial Data*, Wiley, New York, 1991.
12. Briggs, D. J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., Pryl, K., Van Reeuwijk, H., Smallbone, K. and Van Der Veen, A. 'Mapping urban air pollution using GIS: a regression-based approach', *International Journal of Geographical Information Systems*, **11**, 699–718 (1997).
13. Elliott, P., Shaddick, G., Kleinschmidt, I., Jolley, D., Walls, P., Beresford, J. and Grundy, C. 'Cancer incidence near municipal solid waste incinerators in Great Britain', *British Journal of Cancer*, **73**, 702–707 (1996).
14. Whittemore, A. S. and Gong, G. 'Poisson regression with misclassified counts: application to cervical cancer mortality rates', *Applied Statistics*, **40**, 81–93 (1991).
15. Wakefield, J. C. and Morris, S. E. 'Spatial dependence and errors-in-variables in environmental epidemiology', in Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M. (eds), *Proceedings of the Sixth Valencia Meeting on Bayesian Statistics*, Wiley, 1999, pp. 657–684.
16. Carstairs, V. and Morris, R. *Deprivation and Health in Scotland*, Aberdeen University Press, Aberdeen, 1991.
17. Richardson, S. 'Statistical methods for geographical correlation studies', in Elliott, P., Cuzick, J., English, D. and Stern, R. (eds), *Geographical and Environmental Epidemiology*: *Methods for Small-area Studies*, Oxford University Press, Oxford, 1992, pp. 181–204.
18. Richardson, S., Stucker, I. and Hémon, D. 'Comparison of relative risks obtained in ecological and individual studies: some methodological considerations', *International Journal of Epidemiology*, **16**, 111–119 (1987).
19. Piantadosi, S., Byar, D. P. and Green, S. B. 'The ecological fallacy', *American Journal of Epidemiology*, **127**, 893–904 (1988).
20. Greenland, S. and Robins, J. 'Ecological studies: biases, misconceptions and counterexamples', *American Journal of Epidemiology*, **139**, 747–760 (1994).
21. Diggle, P. 'A point process modelling approach to raised incidence of a rare phenomenon in the vicinity of a prespecified point', *Journal of the Royal Statistical Society*, *Series A*, **153**, 340–362 (1990).
22. Diggle, P. and Elliott, P. 'Disease risk near point sources: Statistical analyses for analyses using individually or spatially aggregated data', *Journal of Epidemiology and Community Health*, **49**, S20–S27 (1995).
23. Diggle, P. J. and Rowlingson, B. S. 'A conditional approach to point process modelling of raised incidence', *Journal of the Royal Statistical Society*, *Series A*, **157**, 433–440 (1994).
24. Bithell, J. F. 'An application of density estimation to geographical epidemiology', *Statistics in Medicine*, **9**, 691–701 (1990).
25. Lawson, A. B. and Williams, F. L. R. 'Applications of extraction mapping in environmental epidemiology', *Statistics in Medicine*, **12**, 1249–1258 (1993).

26. Kelsall, J. E. and Diggle, P. J. 'Kernel estimation of relative risk', *Bernoulli*, **1**, 3–16 (1995).
27. Kelsall, J. E. and Diggle, P. J. 'Non-parametric estimation of spatial variation in relative risk', *Statistics in Medicine*, **14**, 2335–2342 (1995).
28. Kelsall, J. E. and Diggle, P. J. 'Spatial variation in risk: a nonparametric binary regression approach', *Applied Statistics*, **47**, 559–573 (1998).
29. Hastie, T. J. and Tibshirani, R. J. *Generalized Additive Models*, Chapman and Hall, London, 1990.
30. Cook-Mozaffari, P. Darby, S., Doll, R., Forman, D., Hermon, C., Pike, M. C. and Vincent, T. 'Geographical variation in mortality from leukaemia and other cancers in England and Wales in relation to proximity to nuclear installations, 1967–78', *British Journal of Cancer*, **59**, 476–485 (1989).
31. Waller, L. A., Turnbull, B. W., Clark, L. C. and Nasca, P. 'Spatial pattern analyses to detect rare disease clusters', in Lange, L., Ryan, L., Billard, L., Brillinger, D., Conquest, L. and Greenhouse, J. (eds), *Case Studies in Biometry*, Wiley, New York, 1994, pp. 3–23.
32. Openshaw, S., Craft, A. W., Charlton, M. and Birch, J. M. 'Investigation of leukaemia clusters by use of a geographical analysis machine', *Lancet*, **i**, 272–273 (1988).
33. Turnbull, B. W., Iwano, E. J., Burnett, W. S., Howe, H. L. and Clark, L. C. 'Monitoring for clusters of disease: application to leukaemia incidence in upstate New York', *American Journal of Epidemiology*, **Supplement 1**, S136–S143 (1990).
34. Besag, J. and Newell, J. 'The detection of clusters in rare diseases', *Journal of the Royal Statistical Society*, *Series A*, **154**, 143–155 (1991).
35. Cuzick, J. and Edwards, R. 'Spatial clustering for inhomogeneous populations', *Journal of the Royal Statistical Society*, *Series B*, **52**, 73–104 (1990).
36. Diggle, P. and Chetwynd, A. 'Second-order analysis of spatial clustering for inhomogeneous populations', *Biometrics*, **47**, 1155–1163 (1991).
37. Schulman, J., Selvin, S. and Merrill, D. W. 'Density equalised map projections: a method for analysing clustering around a fixed point', *Statistics in Medicine*, **7**, 491–505 (1988).
38. Clayton, D. G. and Kaldor, J. 'Empirical Bayes estimates of age-standardized relative risks for use in disease mapping', *Biometrics*, **43**, 671–682 (1987).
39. Eaton, N., Shaddick, G., Dolk, H. and Elliott, P. 'Small-area study of the incidence of neoplasms of the brain and central nervous system among adults in the West Midlands', *British Journal of Cancer*, **75**, 1080–1083 (1997).
40. Besag, J. and Kooperberg, C. 'On conditional and intrinsic autoregressions', *Biometrika*, **82**, 733–746 (1995).
41. Besag, J., York, J. and Mollié, A. 'Bayesian image restoration with two applications in spatial statistics', *Annals of the Institute of Statistics and Mathematics*, **43**, 1–59 (1991).
42. Wolpert, R. L. and Ickstadt, K. 'Poisson/gamma random field models for spatial statistics', *Biometrika*, **85**, 251–267 (1998).
43. Stone, R. 'Investigations of excess environmental risks around putative source: statistical problems and a proposed test', *Statistics in Medicine*, **7**, 649–660 (1988).
44. Morton-Jones, T., Diggle, P. and Elliott, P. 'Investigation of excess environmental risk around putative sources: Stone's test with covariate adjustment', *Statistics in Medicine*, **18**, 189–197 (1999).
45. Lawson, A. 'On the analysis of mortality events associated with a prespecified fixed point', *Journal of the Royal Statistical Society*, *Series A*, **156**, 363–377 (1993).
46. Diggle, P. J., Morris, S. E. Elliott, P. and Shaddick, G. 'Regression modelling of disease risk in relation to point sources', *Journal of the Royal Statistical Society*, *Series A*, **160**, 491–505 (1997).
47. Potthoff, R. F. and Whittinghill, M. 'Testing for homogeneity: I. The binomial and multinomial distributions', *Biometrika*, **53**, 167–182 (1996).
48. Potthoff, R. F. and Whittinghill, M. 'Testing for homogeneity: II. The Poisson distribution', *Biometrika*, **53**, 183–190 (1966).
49. Muirhead, C. R. and Butland, B. K. 'Testing for over-dispersion using an adapted form of the Potthoff-Whittinghill method', in Alexander, F. E. and Boyle, P. (eds), *Methods for Investigating Localized Clustering of Disease*, International Agency for Research on Cancer, 1996, pp. 40–52.
50. Fuller, W. A. *Measurement Error Models*, Wiley, 1987.
51. Carroll, R. J., Ruppert, D. and Stefanski, L. A. *Measurement Error in Nonlinear Models*, Chapman and Hall, London, 1995.
52. Richardson, S. 'Measurement error', in Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (eds), *Markov Chain Monte Carlo in Practice*, Chapman and Hall, New York, 1996, pp. 401–417.

53. Wakefield, J. C. and Stephens, D. A. 'Bayesian errors-in-variables modeling', in Dey, D. K., Ghosh, S. K. and Mallick, B. K. (eds), *Generalized Linear Models*: *A Bayesian Perspective*, Marcel-Dekker, New York, 1999.
54. Jordan, P., Brubacher, D., Tsugane, S., Tsubono, Y., Gey, K. F. and Moser, U. 'Modelling of mortality data from a multi-centre study in Japan by means of Poisson regression with errors in variables', *International Journal of Epidemiology*, **26**, 501–507 (1997).
55. Bernardinelli, L., Pascutto, C., Best, N. G. and Gilks, W. R. 'Disease mapping with errors in covariates', *Statistics in Medicine*, **16**, 741–752 (1997).
56. Diggle, P. J., Tawn, J. A. and Moyeed, R. A. 'Model-based geostatistics (with discussion)', *Applied Statistics*, **47**, 299–350 (1998).
57. Currie, J. 'On the analysis of spatial point process data with inaccurately observed covariate information', in Barnett, V. and Turkman, K. F. (eds), *Statistics for the Environment 4*: *Health and the Environment*, Wiley, 1998.
58. Richardson, S. and Gilks, W. R. 'A Bayesian approach to measurement error problems in epidemiology using conditional independence models', *American Journal of Epidemiology*, **138**, 430–442 (1993).
59. Clayton, D. G. 'Models for the analysis of cohort and case-control studies with inaccurately measured exposures', in Dwyer, J. H., Manning, F., Lippert, P. and Hoflmeister, P. (eds), *Statistical Models for Longitudinal Studies of Health*, Oxford University Press, New York, 1992, pp. 301–331.
60. Spiegelhalter, D. J. 'Bayesian graphical modelling: a case study in monitoring health outcomes', *Applied Statistics*, **47**, 115–133 (1998).
61. Dellaportas, P. and Stephens, D. A. 'Bayesian analysis of errors-in-variables regression models', *Biometrics*, **51**, 1085–1095 (1995).
62. Spiegelman, D. and Casella, M. 'Fully parametric and semiparametric regression models for common events with covariate measurement error, in main study/validation study designs', *Biometrics*, **53**, 395–409 (1997).
63. Doll, R. and Peto, R. *The Causes of Cancer*, Oxford Medical Publications, Oxford, 1981.

